Systematic Classification and Analysis of Themes in Protein−DNA Recognition

Peng Zhou,[†] Feifei Tian,[‡,§] Yanrong Ren,[||] and Zhicai Shang[†,*]

Department of Chemistry, Zhejiang University, Hangzhou 310027, China, College of Bioengineering, Chongqing University, Chongqing 400044, China, Department of Biological and Chemical Engineering, Chongqing Education College, Chongqing 400067, China, and Center for Heterocyclic Compounds, Department of Chemistry, University of Florida, Gainesville, Florida 32611

Received February 1, 2010

Protein−DNA recognition plays a central role in the regulation of gene expression. With the rapidly increasing number of protein−DNA complex structures available at atomic resolution in recent years, a systematic, complete, and intuitive framework to clarify the intrinsic relationship between the global binding modes of these complexes is needed. In this work, we modified, extended, and applied previously defined RNA-recognition themes to describe protein−DNA recognition and used a protocol that incorporates automatic methods into manual inspection to plant a comprehensive classification tree for currently available high-quality protein−DNA structures. Further, a nonredundant (representative) data set consisting of 200 thematically diverse complexes was extracted from the leaves of the classification tree by using a locally sensitive interface comparison algorithm. On the basis of the representative data set, various physical and chemical properties associated with protein−DNA interactions were analyzed using empirical or semiempirical methods. We also examined the individual energetic components involved in protein−DNA interactions and highlighted the importance of conformational entropy, which has been almost completely ignored in previous studies of protein−DNA binding energy.

1. INTRODUCTION

Protein−DNA interactions are involved in the regulation of many important cellular processes, such as transcription, replication, recombination, and translation. Over the past several decades, although great efforts have been made to understand the basic principles that determine specific protein−DNA binding, the molecular mechanism underlying the recognition of DNA by cognate proteins still remains unclear. This is because the structural basis that governs the specific interactions of proteins with DNA is very complicated,[1] and moreover, no simple recognition code between protein amino acids and DNA bases exists.[2]

In recent years, with the rapid progress in the methods and techniques of structural proteomics, use of high-throughput approaches to determine biomolecular structures is becoming facile for experimental chemists and biologists.[3] As a result, the rate of growth of protein−DNA complexes deposited in Protein Data Bank (PDB)[4] has been growing exponentially since 2000. Given the large number of protein−DNA complex structures, how can one explain the specific recognition of particular sequences by regulatory proteins, and how can one predict target sites recognized by proteins? To address these problems, a systematic framework must be constructed to clarify the potential relationships that govern specific protein−DNA binding. Structural classification and systematization of biomolecules and their complexes provide a promising way to achieve this. The best known paradigms of biostructural classifications might be those of Chothia et al.,[5−7] Thornton et al.,[8−10] and Richardson et al.,[11−13] which directly gave rise to the widely used databases SCOP and CATH.[14,15] With the success of SCOP, a number of similar depositories (e.g., SOCR,[16] SCOPPI,[17] and SCOWLP[18]) were launched to systematize the structure and family of biomolecules, and also many efforts (e.g., Pabo and Nekludova,[19] Kim et al.,[20] Ahmad et al.,[21] and Sutch et al.[22]) have addressed comparison and cluster studies of biomolecular interactions.

The structural taxonomy of protein−DNA complexes based on DNA-binding domains of proteins was first described by Harrison[23] and later modified by Luisi.[24] Further, Luscombe et al. presented a comprehensive overview in which the DNA-binding regions in proteins were categorized into eight groups in terms of secondary-structure class and homologous family.[25] Although the DNA-binding-domain-based classification scheme is well-known in the biology community and has been broadly applied to elucidate the structural basis of DNA bound by proteins, it cannot be used to describe the global feature of proteins interacting with multiple sites on DNA sequences, particularly the huge, complicated complexes of DNA with multisubunit proteins such as polymerases, nucleosome core particles, and mismatch repairers. In addition, simply classifying DNA-binding motifs into several types such as HTH, bZIP, and zinc finger[26] is incapable of covering the overall spectrum of various protein−DNA interaction modes. On the other hand, the use of structural descriptors,[27] spatial structure alignment,[28] and specificity scoring[29] to compare binding modes automatically seems to be an alternative approach to

* Corresponding author phone: +86-0571-87952379; fax: +86-0571-87951895; e-mail: shangzc@zju.edu.cn.
[†] Zhejiang University.
[‡] Chongqing University.
[§] University of Florida.
[||] Chongqing Education College.

10.1021/ci100145d
Published on Web 07/23/2010

quantitatively measuring the structural relationship between protein−DNA complexes. However, the classification results obtained merely from machine approaches provide little biological insight and might involve some artificial components.

The concept of themes in protein−nucleic acid recognition first appeared in the perspective article of von Hippel[30] and was later used by Draper to interpret the observed behavior of diverse proteins bound to RNA.[31] By surveying atomic-resolution structures of 20 protein−RNA complexes available at the time, Draper defined two categories of RNA-recognition themes, namely, groove binder and β-sheet pocket. Recently, with a large increase in crystallographic structure data, the two new RNA-recognition themes backbone contactor and packing cavity were suggested to update the old knowledge.[32] The term "theme" could be considered to describe the apparent (global) binding mode of protein to nucleic acid, regardless of the structural characteristics at atom, residue, and sequence levels. Therefore, there is one and only one theme class per protein−DNA complex. In comparison with the nucleic-acid-binding domain, the nucleic-acid-recognizing theme provides a more global, dynamic, and abstract profile to characterize the diverse manners by which proteins interact with specific nucleic acid sequences. In a pioneering work by Jones et al.,[33] protein−DNA interaction footprints were preliminarily classified into three modes, namely, single-headed, double-headed, and enveloping. In the present work, on the basis of the Jones et al.'s scheme, the traditional classification of DNA-binding motifs, and the family and species of DNA-binding proteins collected in SCOP, the previously defined RNA-recognition themes are modified, extended, and applied to describe protein−DNA recognition. First, a classification tree was manually constructed to systematize the DNA-recognition themes observed in currently available high-quality crystal structures of protein−DNA complexes. Subsequently, a representative data set was automatically extracted from the leaves of the classification tree by using an interface comparison algorithm proposed in our recent work.[32] Based on the representative data set, we made a comprehensive investigation on the physical and chemical properties associated with protein−DNA binding. We expect that this work could formulate a foundational and functional framework to clarify the potential relationship between the DNA-recognition themes observed in currently available structural data.

2. MATERIALS AND METHODS

2.1. Setup of Protein−DNA Complex Data Set. In October 2009, there were over 1600 entries containing protein−DNA complexes deposited in the PDB[4] and curated in the Nucleic Acid Database (NDB).[34] These structures were downloaded in Biological Assembly form (rather than the commonly used Asymmetric Unit form) and then examined visually, to avoid including crystal contacts in our complex data set. As a consequence, 1307 eligible entries with a resolution of 3 Å or better were screened out. Further, we used the following criteria to select valid structures for subsequent analysis: (1) Protein complexes with single-stranded DNA molecules were removed from our data set, because single-stranded DNA is more similar to RNA than double-stranded DNA. (2) DNA should have at least five base pairs, because the DNA helix makes one complete turn approximately every 10 base pairs and a definite major/minor groove can only be constituted by at least five base pairs. (3) Protein−DNA complexes were excluded if the DNA had fewer than five nucleotides in contact with the protein. After this procedure, 1192 PDB entries were selected, for which the hydrogen atoms of the proteins and DNA were added and optimized using the REDUCE program[35] and the missing protein side chains were repaired with the SCWRL4 program.[36] Finally, these PDB structures were split into subunits that contained only one DNA helix per subunit, along with all of the contacting protein chains from the parent file. The PDB IDs of the selected and excluded entries are listed in Tables S1 and S2, respectively, of the Supporting Information.

2.2. Assignment of Structural Properties. *Residue−Residue Contact.* We considered two residues (amino acid and/or nucleotide) coming from distinct chains to be in contact (and thus to be interfacial residues) if there was a hydrogen bond, a water-mediated hydrogen bond, a van der Waals interaction, or at least one pair of contacting nonhydrogen atoms ($\leq$6 Å) between them. This "contacting" strategy was modified from the approach of Siggers et al.[28] The hydrogen bonds/water-mediated hydrogen bonds and van der Waals interactions were defined with HBPLUS[37] and PROBE,[38] respectively.

Functional Protein Complex. Assessing the oligomeric state of a protein from its X-ray structure is not always straightforward. Both biological interactions and crystal contacts exist in crystal structures, and there is no clear way to distinguish between them. However, crystal packing interfaces are often smaller than specifically interacting interfaces.[39] This is particularly the case for oligomeric proteins bound to nucleic acids.[40] Therefore, two protein chains were considered to be functionally cooperative if they had at least five residue pairs in contact.

Secondary Structure. Four types of secondary-structure classes, namely, helix, sheet, turn, and loop, were distinguished in this study. The helix class includes the α-helix, π-helix, and 3/10 helix; the sheet class includes the isolated β-strand and multiple β-ladder; the turn class includes 3, 4, and 5 turns defined by a hydrogen bond from C═O (i) to N−H ($i + 3$, $i + 4$, and $i + 5$, respectively); and all others were regarded as belonging to the loop class. The secondary-structure classes of each amino acid residue in proteins were assigned according to the DSSP protocol.[41]

2.3. Comparison of Two Interfacial Geometries. Specific recognition and association in biological systems are closely related to the interfacial architecture of biomolecular complexes, and the geometrical characteristics of biomolecular interfaces can be exploited as a basic parameter to mirror the functional relationship between biological interactions.[42,43] Recently, based on a procedure that dissects the interface geometry in terms of the spatial relationships between individual amino acid−nucleotide pairs, we proposed a new method for quantitatively measuring the geometric similarity between protein−RNA interfaces. In this work, this method was modified slightly to compare interface geometries of protein−DNA complexes sharing a common DNA-recognition theme. Full details of this algorithm can be found in ref 32; a summary is given next.

Briefly, to quantitatively measure the similarity between interfaces in complexes of DNA helices A and B with a

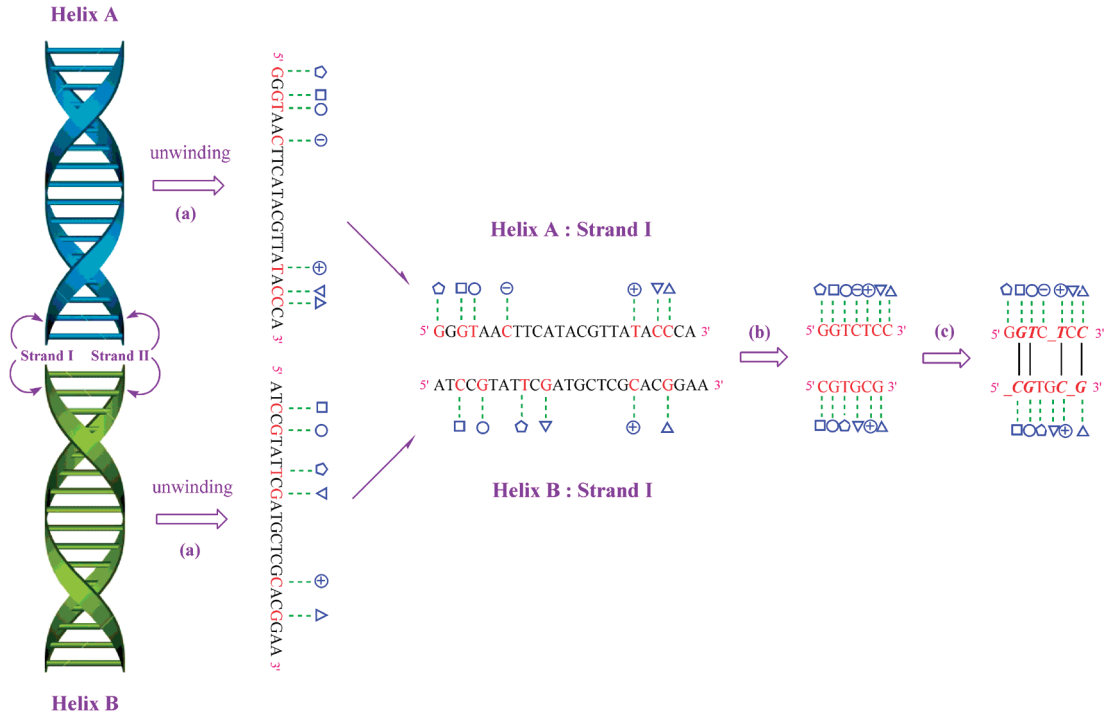

Figure 1. Schematic representation of interface comparison between protein−strand I (helix A) and protein−strand I (helix B) complexes. (Interfacial nucleotides are colored red, and their local geometries are represented in different symbols). (a) Unwinding DNA helices. (b) Removing noninterfacial nucleotides in DNA sequences. (c) Employing dynamic programming to align interfacial nucleotides between two corresponding strands. The interfaces of strand II helices A and B in complex with proteins can be compared in the same way.

protein, we can separately compare the interfaces of corresponding DNA strands in complex with proteins (i.e., strands I and II of helix A correspond to strands I and II, respectively, of helix B) and use average score to quantify the similarity between interfaces in the helix A−protein and helix B−protein complexes. First, we need to identify interfacial nucleotides in the two complexes and determine the spatial locations/directions of contacting amino acids relative to interfacial nucleotides. Then the algorithm proceeds in three steps (Figure 1): (1) unwinding DNA helices, (2) removing noninterfacial nucleotides in DNA sequences, and (3) employing Smith−Waterman dynamic programming[44] to align interfacial nucleotides between two corresponding strands. The element $\mathbf{S}(i, j)$ of the similarity score matrix $\mathbf{S}$ used in dynamic programming is a measure of local geometry similarity between the interfacial nucleotide i in strand I (or II) of helix A and the interfacial nucleotide j in strand I (or II) of helix B. (The local geometry of a nucleotide at the interface of a protein−DNA complex is defined as the spatial distribution of neighboring amino acid residues relative to this nucleotide.) The comparison results in a quantified value called the interface similarity score (ISS) which falls into the range from 0 to 100; 100 indicates identical geometries for the two complex interfaces, and 0 corresponds to the other extreme where the two interfaces are completely independent of each other. As can be seen, this comparison algorithm is locally sensitive because it handles the interfacial geometry at individual residue level and employs Smith−Waterman algorithm (i.e., the local alignment algorithm) to perform nucleotide matching.

It worth noting that, although this interface comparison method is based on the primary sequence of the compared DNA pair in complex with their protein partners, the resulting ISS score could be used as an effective measure of spatial geometric similarity between the protein−DNA interfaces. This is true for the following reasons: (1) Protein−DNA complexes with similar interface geometries normally share similar primary sequences because of their biological homologousness, whereas the ISS scores of those with distinct interface geometries must not be too high, although our method might miss some possibilities of the ordering manners of the strings. (2) More importantly, the regular DNA double-helix structures are essentially different from those of complicated proteins; DNA structures retain an almost-linear state without wrapping and folding, so their spatial arrangements basically agree well with their primary sequences.

2.4. Statistics of Amino Acid−Nucleotide Doublet Propensities. The doublet propensities for different types of amino acid−nucleotide pairs can be statistically derived from the protein−DNA complex structures using the function[45]

$$P_{ab} = \frac{N_{ab}/\sum N_{ij}}{N_a N_b/(\sum N_i)(\sum N_j)}$$

where N_{ab} and $\sum N_{ij}$ are the number of pairs of amino acid a in contact with nucleotide b and the total number of any amino acid in contact with any nucleotide in the data set, respectively; N_a and N_b are the total numbers of amino acid a and nucleotide b, respectively, in the data set; and $\sum N_i$ and $\sum N_j$ are the total numbers of all amino acids and all nucleotides, respectively. A propensity greater than 1 indicates that a given amino acid occurs more frequently in the protein−DNA interface with a given nucleotide than in the remainder, whereas a propensity less than 1 indicates that a given amino acid occurs less frequently in the interface with a given nucleotide.

2.5. Calculation of Energetic Components Involved in Protein−DNA Binding. On the basis of the systematic application of the AMBER99 force field[46] to reproduce the experimentally determined interaction energies of 30 protein−DNA complexes and the interaction energy changes involved in mutations of 189 protein−DNA complexes, Donald et al. suggested that nonbonding potentials might perform better before than after structure minimization.[47] In addition, we have also examined the effect of molecular mechanics minimization on the calculated energies of several complex structures and found that there was no significant difference between the calculated values derived before and after the minimization. Therefore, the protein−DNA complex structures used here for energetic analysis were not optimized. The noncovalent contribution to the free energy change associated with protein−DNA binding were decomposed into five individual components, namely, hydrogen bonding, van der Waals contact, desolvation effects, electrostatic interaction, and conformational entropy loss, that were evaluated through empirical or semiempirical approaches as described in the following subsections. We did not consider the effect of water molecules explicitly in this study; the contribution of water is implicitly involved in the desolvation term.

Hydrogen Bonding. A short X−H···Y (X = N or O, Y = N or O) contact across protein−DNA interface was identified as a hydrogen bond if it satisfied the geometrical constraint defined in HBPLUS.[37] Then, its dissociation energy was estimated with the angle-weighted Lennard-Jones 8−6 potential[48]

$$\Delta G_{\mathrm{hb}} = \varepsilon_{ij}\left[3\left(\frac{d_{ij}^{*}}{d_{ij}}\right)^{8} - 4\left(\frac{d_{ij}^{*}}{d_{ij}}\right)^{6}\right]\cos^{4}\theta$$

where θ is the angle $\angle$(X−H···Y); d_{ij} is the separation between the hydrogen-bond donor i (X) and the hydrogen-bond acceptor j (Y); and ε_{ij} is the optimum hydrogen-bond energy for the particular hydrogen-bonded atoms i and j, considering that d_{ij}^{*} is the optimum hydrogen-bond length. ε_{ij} and d_{ij}^{*} vary according to the chemical type of the hydrogen-bonded atoms i and j, for which Boobbyer et al. assumed $\varepsilon_{ij} = 2.0$ kcal·mol^{-1} and $d_{ij}^{*} = 3.2$ Å for N−N hydrogen bonds, $\varepsilon_{ij} = 2.8$ kcal·mol^{-1} and $d_{ij}^{*} = 3.0$ Å for N−O hydrogen bonds, and $\varepsilon_{ij} = 4.0$ kcal·mol^{-1} and $d_{ij}^{*} = 2.8$ Å for O−O hydrogen bonds.[48]

van der Waals (vdW) Contact. The vdW contacts at the protein−DNA interface were detected with PROBE,[38] and their contacting energies were typically expressed with the Lennard-Jones 12−6 potential[46]

$$\Delta G_{\mathrm{vdW}} = E_{ij}\left[\left(\frac{D_{ij}^{*}}{d_{ij}}\right)^{12} - 2\left(\frac{D_{ij}^{*}}{d_{ij}}\right)^{6}\right]$$

where E_{ij} is the Lennard-Jones well depth and D_{ij}^{*} is the distance at the Lennard-Jones minimum. The Lennard-Jones parameters between pairs of different atom types were obtained from the Lorentz−Berthelodt combination rule,[49] and the atomic parameters were taken from the AMBER99 force field.[46]

Desolvation Effects. The free energy change accounting for the desolvation accompanying protein−DNA binding was estimated (assuming that the conformation of interfacial residues is invariant upon complex formation) as[50]

$$\Delta G_{\mathrm{desolv}} = \Delta\mathrm{SASA}^{\mathrm{phob}}\sigma^{\mathrm{phob}} - \Delta\mathrm{SASA}^{\mathrm{phil}}\sigma^{\mathrm{phil}}$$

where $-\Delta$SASA$^{\mathrm{phob}}$ and $-\Delta$SASA$^{\mathrm{phil}}$ are, respectively, the hydrophobic and hydrophilic solvent-accessible surface areas (SASAs) buried upon binding and σ^{phob} and σ^{phil} (with values of 0.012 and −0.06 kcal·mol^{-1}·Å^{2}, respectively) are the hydrophobic and hydrophilic atomic solvation parameters (ASPs) obtained by fitting experimentally measured transfer free energies.[51] We computed atomic SASAs using Sanner et al.'s algorithm implemented in MSMS[52] with the ProtOr radius (for protein atoms/groups),[53] extended ProtOr radius (for DNA atoms/groups),[54] and 1.4-Å radius (for the water probe).

Electrostatic Interaction. The contribution of electrostatic interactions between a protein and a DNA molecule to the binding free energy was calculated as

$$\Delta G_{\mathrm{ele}} = G_{\mathrm{ele}}^{\mathrm{protein-DNA}} - G_{\mathrm{ele}}^{\mathrm{protein}} - G_{\mathrm{ele}}^{\mathrm{DNA}}$$

where G_{ele} is the total electrostatic energy of a solute molecule and includes two terms, one for the polar solvation energy and one for the Coulombic interaction energy, which is the work of charging the solute molecule in the Poisson−Boltzmann (PB) equation.[55] This method approximates the electrostatic solvation free energy as the reaction field energy of removing a solute molecule from a vacuum to water. In this study, the finite-difference solution of the PB equation was implemented in DelPhi[56] with a temperature of 298.15 K, an ionic strength of 0.15 M, and dielectric constants of 4 for the solute and 80 for the solvent. A grid spacing of 2.0 grid/Å, in which the longest linear dimension of the solute occupied 80% of the lattice, was used to determine the size of the cubic lattice,[57] and the boundary potentials were set to the sum of the Debye−Hückel values.[58] Instead of the sophisticated solvent-exclusion (SE) surface, the van der Waals (vdW) surface was used here to serve as the dielectric boundary, because Qin and Zhou recently demonstrated that the vdW dielectric boundary in PB calculations could give a more reasonable result for the electrostatic contribution to protein−nucleic acid binding.[59] The PARSE parameter set[60] was used to assign partial charges and radii for the atoms in studied protein−DNA complex systems.

Side-Chain Conformational Entropy Loss. The rotatable side chains of interfacial amino acid residues are frozen as a result of protein−DNA binding, leading to a significant loss in conformational entropy (and thus in conformational energy).[61] For an interfacial residue i, its side-chain conformational entropy (SCE) can be calculated simply by virtue of the Boltzmann formulation as[62]

$$S_i = -R\sum_j p_i^j \ln p_i^j$$

where R is the universal gas constant, 1.987 cal·mol^{-1}·K^{-1}; p_i^j is the probability of residue i being in state j; and the sum is taken over all side-chain rotamer states of the residue i. The rotamer library used here was developed by Lovell et al.,[63] as recommended by Dunbrack for the study of SCE,[64] and the probability distribution of side-chain rotamers was derived from the Boltzmann distribution with the classical Lennard-Jones potential.[65] The side-chain conformational

energy during the binding is therefore defined as $\Delta G_{SCE} = -T\sum_i \Delta S_i$, where the sum is over all interfacial residues and T is room temperature, 298.15 K. This procedure was implemented with the in-house program 2D-GraLab.[66]

In addition to energetic components, the interface area B was also considered to be closely related to the thermodynamic properties of protein−DNA binding and was used as a measure to account for the loss in protein and DNA surfaces that are buried upon contact between the two macromolecules. According to Nadassy et al.'s suggestion,[67] B was defined as the area of the surface buried on both the protein and the DNA during the binding, that is, $-\Delta$SASA. Similarly, the hydrophobic and hydrophilic counterparts, B^{phob} and B^{phil}, can be defined to $-\Delta SASA^{phob}$ and $-\Delta SASA^{phil}$, respectively.

3. RESULTS AND DISCUSSION

3.1. Classification Tree. It is a difficult task to classify the DNA-recognition themes in a definite framework, because nearly all DNA-binding proteins have structural similarities with each other and, in some of these cases, share a common evolutionary origin. Therefore, we used a strategy inspired by the SCOP construction (which is fulfilled by manual inspection in conjunction with automated methods)[14] to plant the classification scheme of DNA-binding themes. First, the complex structures were examined one by one to distinguish between the apparently different binding modes, irrespective of the sequence pattern and the fluctuation in local conformation. Then, the DNA-binding domains of complexes with apparently similar binding modes were further compared in pairs using the structure alignment programs SSAP[68] and ProFit.[69] The families and species of all complexes were also inspected to determine the potential differences among their binding modes.

Through this procedure, a systematic and comprehensive classification tree was designed to intuitively describe the theme relationship between all protein−DNA binding modes observed in our data set. This tree grows from a root, through trunks and branches, to leaves (Figure 2). At the trunk level, DNA-recognition themes are classified as groove binder, multisite recognizer, or cross-section contactor. A groove binder places one or more substructures (helix, strand, or loop) in the major and/or minor grooves of DNA helix, recognizing both the specific sequence of bases and the shape or dimension of the groove (Figure 3a); a multisite recognizer uses a complex combination of spatially dispersed segments to create a binding surface or pocket that examines the specific arrangement of DNA bases (Figure 3b); and a cross-section contactor provides a small region on its surface to nonspecifically touch the transversal of DNA helix (Figure 3c). The groove-binder and cross-section-contactor binding modes with DNA are relatively simple, so these two themes are further categorized according to the secondary-structure assignment of protein motifs that directly interact with DNA, whereas the more complicated multisite-recognizer theme is divided into ruleless contact, backbone contact, and β-sheet pocket in terms of binding behavior. The further classifications of these DNA-binding themes are based on the arrangement of secondary-structure elements, the traditional definition of DNA-binding motifs, the protein family and species, and so on. Eventually, each leaf of the tree consists of several basic themes that cannot be further divided because of the limited structure data available currently. Each basic theme contains one or more PDB entries of which the protein−DNA complexes share a consensus binding mode with ISS scores, in pairs, of more than 80. In this way, a total of 200 basic themes were defined for current PDB depositions; their members and representatives are summarized in Table S3 of the Supporting Information. Detailed descriptions of each trunk, branch, and leaf in the classification tree are also provided in Table S4 of the Supporting Information. A previous examination of genes that are functionally assigned in the PEDANT database[70] showed that typically 2−3% of a prokaryotic genome and 6−7% of a eukaryotic genome encodes DNA-binding proteins.[26] Therefore, the classification tree presented here is far from complete and, thus, is not exclusive and invariable. For example, a leaf might be upgraded to a branch when additional crystallographic data are available in the future. It should also be noted that the number of structures in the PDB does not necessarily reflect the relative importance of the protein−DNA complexes in the organism. This means that the member sizes of different themes do not always coincide with the usage frequency of these themes in the entire protein−DNA complex pool.

3.2. Thematically Nonredundant Data Set. Each leaf of the classification tree consists of one or more basic theme types, and each basic theme type includes one or more PDB entries containing protein−DNA complexes sharing the common basic theme. Hence, we can select a representative for each basic theme and use these representatives to comprise a thematically nonredundant data set—each basic theme occurs only once in this data set. Apparently, the structural/geometrical features are in close agreement between the interfaces of protein−DNA complexes that belong to the same basic theme. Therefore, the locally sensitive algorithm is a good choice for comparing interface geometries of protein−DNA complexes sharing a common basic theme. Here, we used the following strategy to select a representative for a basic theme: (1) For a theme consisting of single member, this one member was selected as representative. (2) For a theme consisting of two members, the one with higher resolution was selected. (3) For a theme consisting of three or more members, the average ISS values between each member and the remainder were calculated in turn, and the one with the highest average ISS was selected. In this way, a nonredundant data set containing 200 thematically diverse protein−DNA complexes was built for the purpose of studying protein−DNA interaction and recognition. This data set is summarized in Table 1 and its details are provided as Table S3 in the Supporting Information.

Previously, representative data sets for protein−protein and protein−RNA interactions have been developed by different groups[20,32,43,71−74] and successfully applied to estimate the total number of protein folds and interactions,[75,76] compare singular protein folds and interface folds,[43,76] and predict protein interactions,[77−80] among other uses. Therefore, the newly defined data set for protein−DNA interaction constitutes an important step in this direction and is expected to be used in these fields as well. The analysis and discussion in the following sections are all based on this representative data set.

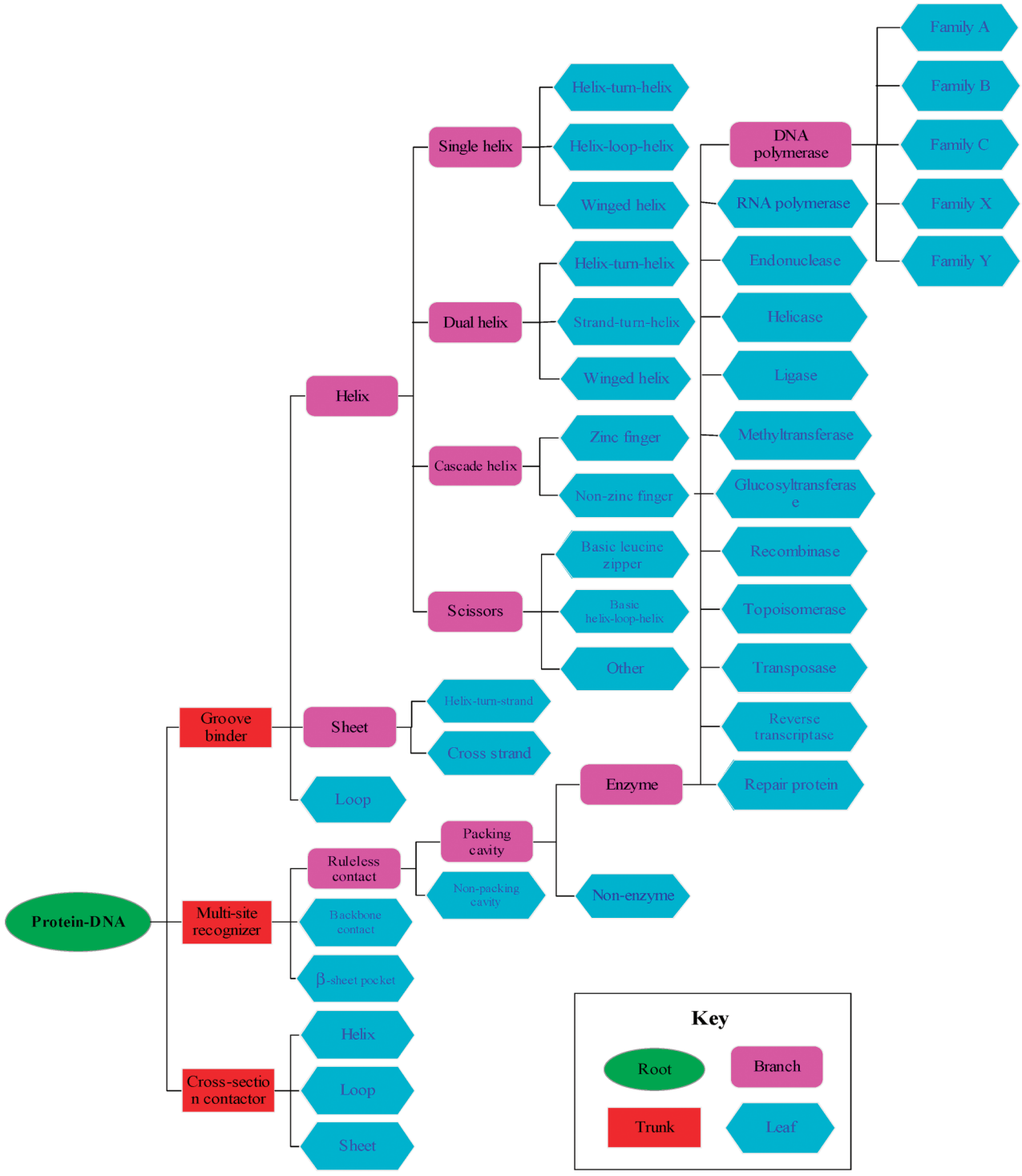

Figure 2. Classification tree for DNA-recognizing themes. This tree is constructed just for the protein−DNA complexes currently available from the PDB; hence, it is far from complete and is not exclusive and invariable. The goal of designing this tree was to provide an intuitive, systematic, and comprehensive framework to clarify the potential relationship between the global binding modes of structurally known protein−DNA complexes. Detailed descriptions of each trunk, branch, and leaf in the tree are provided in Table S4 of the Supporting Information.

3.3. Amino Acid−Nucleotide Doublet Propensity. Statistical analysis of residue contacting propensities at biomolecular interfaces provides straightforward insight into the structural basis of molecular recognition and is used to explore questions such as how the apparent contact codes govern the binding that occurs. Based on the newly defined representative data set, we first examined the amino acid−nucleotide doublet propensities in protein−DNA complexes; the statistical results are shown in Figure 4a. It is expected that positively charged amino acids such as Arg and Lys and polar amino acids such as Asn and Ser are preferable for pairing with nucleotides, because these amino acids are electrostatically complementary to the DNA backbones and/or can specifically hydrogen bond to bases. In contrast, negatively charged amino acids such as Glu and Asp and nonpolar bulk amino acids such as Leu and Phe exhibit a markedly reverse behavior in interaction with nucleotides. Interestingly, the neutral Gly has a higher

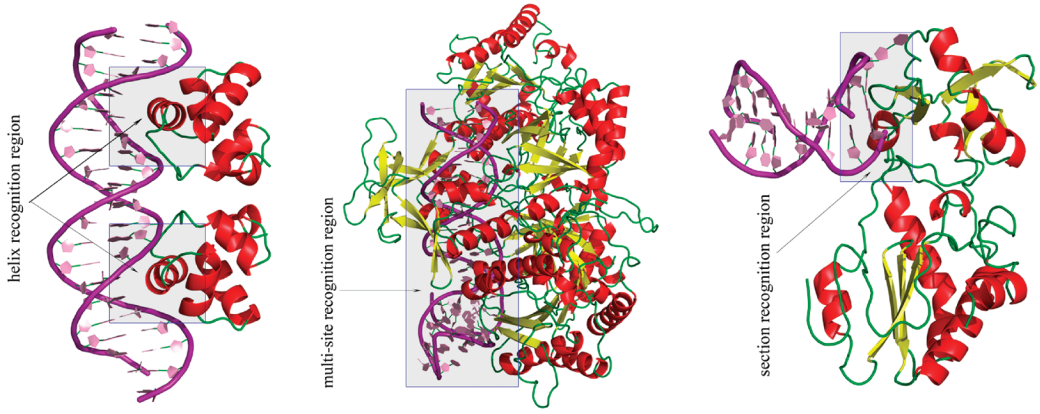

Figure 3. Some examples of DNA-recognizing themes at the trunk level. (a) Groove binder. The phage 434 repressor dimer places its two helices in the major grooves of a DNA segment (PDB: 1PER). (b) Multisite recognizer. The Ku heterodimer uses an irregular pocket to accommodate a DNA segment (PDB: 1JEY). (c) Cross-section contactor. The N-terminal fragment of leukemia virus reverse transcriptase presents two antiparallel strands on its surface to touch with the transversal of a DNA segment (PDB: 1D0E).

Table 1. Thematically Nonredundant Data Set[a]

groove binder (76)	**2HDD (21)**, 1JKO (9), 2O4A (7), 1J1V (4), 3GNA (2), 1DSZ (12), 1KB4 (5), 1W0U (3), 3DFX (1), 1SKN (2), 1VAS (2), 1BC8 (9), 2PI0 (4), 2C6Y (9), 1RH6 (3), 2P5L (1), 1RPE (5), **2PUD (24)**, 1HLV (1), 1QPI (3), 1LLI (3), 1ZS4 (1), 3CLC (1), 1IC8 (1), 1LE8 (3), 3FDQ (1), 1ZG1 (3), 6PAX (4), 2E1C (1), 1IGN (1), 1BL0 (2), 1LQ1 (1), 1H6F (2), 1AIS_B (2), 1L3L (2), 1TRO (2), 1JJ4 (2), 1B3T (1), 1XSD (2), 1J59 (14), 2ISZ (5), 3D6Z (8), 3GZ6 (1), 2UZK (1), 1Z9C (1), 2NNY (1), 3H0D (2), 1REP (1), 1F4K (2), 6CRO (1), **3G9P (20)**, **1A1L (21)**, 1H89_C (3), 1CF7 (1), 3HTS (1), 1CKT (3), 2Z3X (1), 3D1N (1), **2E42 (21)**, 1AM9 (9), 1GDT (1), 2W7N (1), 1EWN (4), 3ORC (1), 3C2I (1), 1ODH (1), 1DP7 (1), 1OZJ (2), 1A73 (6), 2EZV (1), 1MJM (17), 1HAO (3), 1H9D (3), 2ITL (3), 3GFI (1), 1PH6 (10)
multisite recognizer (114)	3CWS (9), 2IS6 (5), 2QNF (1), 2FIO (1), 1EGW (4), 2C5R (1), 1FZP (1), 1SFU (9), 1OE4 (3), 3BIE (8), 3HQF (1), 2WIW (1), **1T9I (29)**, 1C8C (18), **1QN9 (24)**, 1OWF (10), **1NK8 (48)**, 1X9W (16), 2PYJ (3), 2OZS (14), 3IAY (1), 2VWJ (1), 3F2B (3), **1ZJM (95)**, **2BR0 (83)**, 2AQ4 (3), 3C46 (3), 3BAM (4), 1DMU (1), 1DFM (2), 1DC1 (1), 2P0J (2), 2GB7 (3), 1WTE (1), 1QRH (7), 3HQG (1), **1EON (27)**, 1FOK (1), 1TX3 (18), 2E52 (1), 2FKC (3), 3GOX (2), 1SA3 (2), 2ODI (3), 1IAW (1), 1FIU (1), 3C25 (1), 2PVI (5), 3DPG (3), 1CW0 (2), 2P6R (1), 2Q2T (2), 2OWO (1), 1X9N (1), 1DCT (1), 2JG3 (9), 1Y8Z (4), 1M5R (4), 1P4E (3), **1F44 (22)**, 2A3V (1), 1A31 (11), 2H7G (2), 2RGR (1), 1MUS (5), 1R0A (5), 2AOR (2), 1OH6 (14), 1RRQ (3), 1JEY (1), 1NFK (18), 2I06 (4), 3BEP (1), 1OMH (6), 1BG1 (2), 1U8B (1), 3BS1 (1), 2DNJ (5), 3GX4 (2), 2VLA (1), 1PT3 (3), 1V15 (2), 1TEZ (1), 1YFJ (3), 1ORP (3), 1QUM (3), 2W36 (2), **1PM5 (27)**, 2VE9 (1), 3COQ (8), **2C7Q (25)**, 3EI2 (2), 2VIH (6), 1I3J (3), 1R71 (1), 3BRG (4), 2VY1 (2), 3DSD (2), 1D02 (1), 2EUX (10), **1P3L (29)**, **2NOH (25)**, 2AHI (10), 1PDN (1), 2QSH (1), 3CVU (4), 2QKB (4), 1LRR (4), 1Z63 (1), 1CEZ (4), 2OFI (1), 1SSP (12), 2ZO1 (9), 1OUP (2)
cross-section contactor (10)	1UUT (1), 2W42 (1), 2IHN (1), 3H25 (1), 2BGW (1), 2R9L (1), **1ZTT (20)**, 1RXW (1), 1MTL (1), 2VOA (1)

[a] In a PDB entry, all of the protein chains in contact with DNA were considered if no chains are specified by chain names following the PDB id, separated by a underscode. The number of members belonging to the representative's theme is given in parentheses. The representatives with members more than or equal to 20 are shown in bold.

propensity for contact with DNA, which could be explained by the fact that the small, flexible Gly is readily engaged in DNA-binding motifs to serve as linkers between non-neighboring recognition fingers but without introducing additional penalties such as distortions and clashes. On the other hand, there is no significant difference between the propensities of the four nucleotides interacting with proteins, and only guanine has a slightly higher value than the other three nucleotides. These findings, except that for Gly, were compatible with previous observations on protein−nucleic acid interactions.[81−84]

For the purpose of comparison, the amino acid−nucleotide doublet propensities in the entire crude protein−DNA complex pool (consisting of 1192 PDB entries) were also inspected. These are shown in Figure 4b. Upon comparing parts b and a of Figure 4, one can see that the doublet

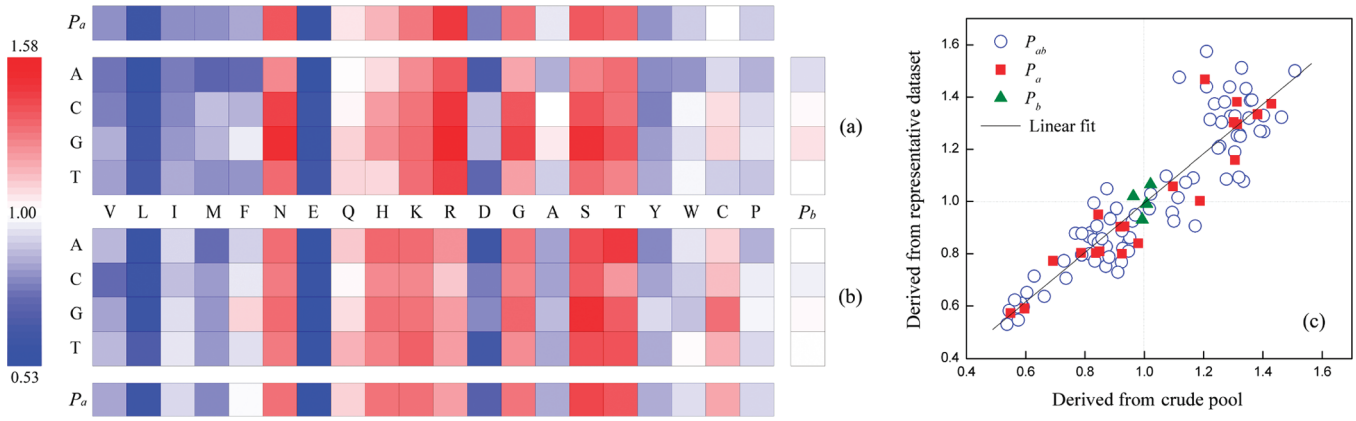

Figure 4. Graphical matrices show the amino acid−nucleotide doublet propensities P_{ab}, P_a, and P_b derived from (a) the representative data set and (b) the crude pool. P_a and P_b denote the propensities of each amino acid coming into contact with all four nucleotides and each nucleotide coming into contact with all 20 amino acids, respectively. The colors indicate whether a particular doublet was observed more (red) or less (blue) than expected (white). Higher color intensity indicates observations more extreme from the expected value than colors of lower intensity. (c) Scatter plot showing the correlation between the doublet propensities derived from the representative data set and the crude pool, with correlation coefficient $R = 0.921$.

Table 2. Mean Values and Standard Deviations of Interface Areas ($\bar{B}$, $\bar{B}^{phil}$, and $\bar{B}^{phob}$) and Energetic Components ($\Delta\bar{G}$) Derived from the Representative Data Set

theme	number	$\bar{B}$(Å^2)	$\bar{B}^{phil}$(Å^2)	$\bar{B}^{phob}$(Å^2)	$\Delta\bar{G}_{HB}$ (kcal·mol^{-1})	$\Delta\bar{G}_{vdW}$ (kcal·mol^{-1})	$\Delta\bar{G}_{ele}$ (kcal·mol^{-1})	$\Delta\bar{G}_{desolv}$ (kcal·mol^{-1})	$\Delta\bar{G}_{SCE}$ (kcal·mol^{-1})
groove binder	76	2446.1 ± 1078.0	1339.9 ± 559.0	1106.5 ± 560.2	−24.4 ± 17.3	−5.0 ± 4.2	−16.1 ± 12.2	10.4 ± 8.6	6.2 ± 3.6
multisite recognizer	114	3338.4 ± 2470.2	1807.6 ± 1403.7	1530.8 ± 1100.7	−34.3 ± 31.9	−8.5 ± 6.7	−21.1 ± 25.6	10.7 ± 14.7	7.8 ± 8.3
cross-section contactor	10	1729.3 ± 1007.3	907.3 ± 553.2	822.0 ± 458.7	−11.8 ± 5.8	−4.9 ± 2.6	−7.7 ± 6.6	2.4 ± 2.8	3.4 ± 2.1
all	200	2602.8 ± 2050.8	1430.5 ± 1151.1	1172.3 ± 933.2	−26.2 ± 27.0	−5.0 ± 5.9	−16.8 ± 21.0	11.5 ± 12.4	6.7 ± 6.8

propensities derived from the crude pool are quite similar to those obtained from the representative data set, with a high correlation of $R = 0.921$ between them (Figure 4c); the slight difference can be ascribed to the additional effects arising from the redundance of the crude pool. This comparison demonstrates that the representative data set presented here is a good reflection of all of the structure-known protein−DNA complexes retrieved from the PDB and could serve as a reliable source for studying the molecular mechanism underlying protein−DNA recognition and binding.

3.4. Energetic Components Involved in Protein−DNA Binding. In protein−DNA systems, the intermolecular association can roughly be framed by dividing the interaction into two energetic components: direct and indirect readouts.[85] The indirect readout arises from the conformational changes (e.g., bending) in DNA when it is bound by protein. In this study, we focused on the direct readout, which refers to the specific noncovalent interactions (e.g., hydrogen bonds) between protein amino acids and DNA nucleotides. The direct binding energy of a protein−DNA complex can be predicted quantitatively using the knowledge-based potential and empirical force field approach.[47] However, here, we are interested only in analysis (rather than prediction) of individual energetic components involved in protein−DNA binding.[86,87] Therefore, we separately investigated five noncovalent interactions contributing to the binding: hydrogen bonding, van der Waals contact, electrostatic interaction, desolvation effects, and conformational entropy loss. In addition to these energetic components, the interface area of protein−DNA complexes was also examined because this property is closely related to the thermodynamic processes accompanied with biomolecular recognition. The statistics for mean interface areas (including total, hydrophobic, and hydrophilic) and mean energetic contributions derived from the representative data set are summarized in Table 2; a detailed list of interface areas and individual energetic contributions for the 200 representatives is provided in Table S5 of the Supporting Information.

Interfacial Area and Its Correlations with Individual Energetic Contributions. La Conte et al. found that most protein−protein interface areas are in the range of 1200−2000 Å^2.[88] The protein−DNA interface area sizes, however, seem to be larger than those buried in protein−protein association, with most varying from 1500 to 6000 Å^2 and a few even reaching over 15000 Å^2. The large average interface area of protein−DNA binding is mainly due to the multisite-recognizer theme, where the proteins provide an irregular pocket that effectively wraps around the DNA. Another difference between protein−DNA and protein−protein interfaces is that the former are relatively hydrophilic, whereas the latter are hydrophobic, as in the protein core:[89] for almost all protein−DNA interfaces, the hydrophilic area, B^{phil}, is always larger than the hydrophobic area, B^{phob}.

The distribution range of interface areas within a leaf is very narrow. This is anticipated because the members belonging to the same leaves must be quite similar in their interface sizes. In contrast, the distribution state of interface areas at the trunk level varies more significantly, and the groove-binder and cross-section-contactor themes have relatively small areas as compared to those of the multisite-recognizer theme. This is reasonable because the multisite-recognizer theme includes many huge, complicated protein−DNA complexes such as polymerases, Holliday junctions, and mismatch repairers. The scatters of five individual energetic contributions (ΔG) against total interface area (B) for the 200 representative complexes are plotted in Figure 5. It can be seen that there is a good correlation between these energetic contributions and the interface area, with the

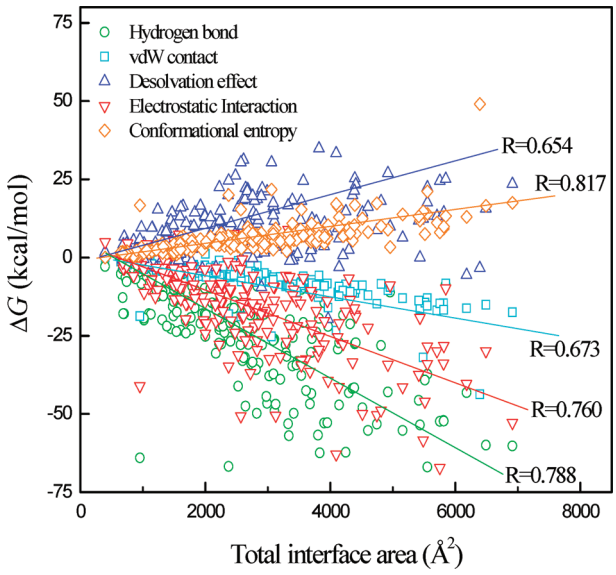

Figure 5. Scatters of five individual energetic contributions versus the total interface area for the 200 representatives.

correlation coefficients R all above 0.65. That is, the larger the interface is, the greater the energetic contributions are. This is expected because a larger interface allows for more nonbonding interactions. Among the five energetic components, hydrogen bonding, van der Waals contact, and electrostatic interaction are negatively correlated with the interface area, because of their favorable contribution to binding ($\Delta G < 0$), whereas desolvation and conformational entropy exhibit a positive correlation with the interface area, given that they introduce an energetic penalty to the binding ($\Delta G > 0$).

Hydrogen Bonding. Hydrogen bonding is traditionally believed to confer the most specificity for protein−DNA recognition. Herein, we focused on the role of hydrogen bonds in maintaining the stability of protein−DNA complexes. The results from a statistical analysis showed that the average number of hydrogen bonds is 20.95 per interface or one per 124.33 Å^2 of interface area. There is a difference in the average number of intermolecular hydrogen bonds between protein−DNA and protein−protein interfaces; the latter exhibit only ∼10 hydrogen bonds per interface.[88] The numbers of hydrogen bonds that we found at the representative protein−DNA interfaces were comparable to those reported in previous work.[90] The large numbers of hydrogen bonds observed suggest a significant energetic contribution to the binding, which can be rationalized by the -26.24 ± 27.0 kcal·mol^{-1} dissociation energy of hydrogen bonds per interface.

van der Waals Contact. Even though van der Waals (vdW) contacts are reasonably weak, the ubiquity of interatomic contacts in/at tightly packed biomolecular interiors/interfaces implies that the role of vdW forces cannot be neglected in shaping macromolecular architecture. An early study found that dispersion forces, with an r^{-6} dependence, provide an additional 7 kcal·mol^{-1} for the stabilization of trypsin−trypsin inhibitor (protein−protein) complex,[91] and later, Roth et al. gave a detailed discussion of the theoretical models of vdW interactions involved in proteins.[92] Although some studies have shed light on the importance of vdW forces in protein−DNA interactions,[47,93] there has not been a systematic study on the vdW contribution to protein−DNA binding. Herein, we tried to conduct a preliminary investigation of this topic. To dissect the atomic structure of protein−DNA binding sites, the Voronoia approach[94] was employed to analyze the packing densities (PDs) of interfacial and noninterfacial atoms, where the packing density is defined as the ratio of the atomic vdW volume to the sum of the atomic vdW volume plus the solvent-excluded volume. The average PD values of interfacial DNA atoms, interfacial protein atoms, noninterfacial DNA atoms, and noninterfacial protein atoms were found to be 0.52, 0.54, 0.57, and 0.68, respectively, indicating that the interfacial atoms, whether they belong to protein or to DNA, have similar degrees of packing[54] and the noninterfacial atoms, especially the protein atoms, are more likely to be closely packed than the interfacial ones (this is mainly due to the atoms in the tightly packed core of proteins). Further, the vdW packing energy at the interface was calculated for each representative complex using the Lennard-Jones potential in conjunction with the AMBER99 parameter set.[46] The packing energies of the groove-binder, multisite-recognizer, and cross-section-contactor themes and all representatives are -4.96 ± 4.2, -8.47 ± 6.7, -4.88 ± 2.6, and -5.04 ± 5.9 kcal·mol^{-1}, respectively. It is evident that the packing energy relates to the interface size; compared with the groove-binder and cross-section-contactor themes, the multisite-recognizer theme has the largest interface and also the strongest packing energy. This can be confirmed by plotting the quantitative correlation between the packing energy and interface area. As shown in Figure 5, most scattering points (cyan squares) are closely distributed along the fitting line, but there are also a few outliers that deviate significantly from the line. This phenomenon is not unexpected because the short-range vdW potential is very susceptible to the interatomic distance and errors involved in structural data (particularly those with lower resolutions, e.g., 3 Å) can sometimes give rise to an abnormal result for the calculated energy. In addition, it is worth noting that the polar nature of protein−DNA packing interfaces can lead to a slight shrinkage of the atomic vdW radii, resulting in the so-called Coulombic radii.[95] The shorter atomic radii might give better compatibility in the local region of the packing interfaces.

Electrostatic Interaction. Electrostatic complementarity is known to be the central chemical force giving rise to protein−DNA association. However, many theoretical calculations have found that electrostatic interaction energies between proteins and nucleic acids are positive, meaning that electrostatic interactions are destabilizing for their binding.[96−98] There are two opposite electrostatic effects of binding, namely, unfavorable desolvation penalties incurred by buried polar groups and favorable bridge energy due to the Coulombic interactions; the dominant effect can determine the final electrostatic role of stabilization or destabilization. According to our calculations, in most of the representative complexes, the absolute values of the desolvation penalty and bridge energy are quite significant (both accounting for more than 50 kcal·mol^{-1}), and the stabilization appears to overpower the destabilization; the electrostatic interaction energy between proteins and DNA is -16.79 ± 21.0 kcal·mol^{-1}, indicating a favorable contribution to the protein−DNA binding. Our findings were consistent with those of Qin et al., who launched a comparative study to explore the electrostatic nature of protein−RNA systems.[59]

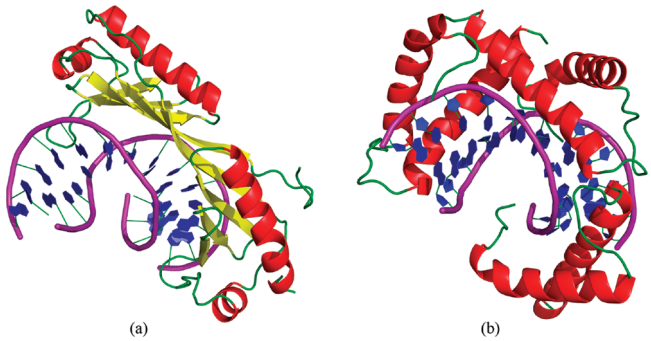

Figure 6. Two typical examples showing protein−DNA interactions dominated by hydrophobic forces. The proteins (a) TBP and (b) SASP place β-sheet and α-helix elements, respectively, in DNA minor grooves to closely contact the aromatic bases, but the negatively charged DNA backbones remain untouched from the protein moieties (PDB: 1QN9 and 2Z3X). It is noteworthy that the DNA sequence compositions in the two complexes are quite different: one is TA-rich (a), and the other is CG-rich (b).

Subsequently, the electrostatic energy was fitted separately to the total interface area B (Figure 5), the hydrophilic area B^{phil}, and the hydrophobic area B^{phob}. Unexpectedly, the three fitting correlation coefficients R are very close (around 0.75). This is contrary to the expectation that the electrostatic energy should be more correlated with the hydrophilic (polar) area than with the total and hydrophobic (nonpolar) areas. By data analysis, we found that the ratios among the areas, $B/B^{\text{phil}}/B^{\text{phob}}$, for many representatives are roughly a constant, which could explain the similar correlations between the electrostatic energy and the different interface areas. In addition, the stable ratio found here implies a conservative interface composition among different protein−DNA complexes.

Desolvation Effects. The solvation free energy of transferring a molecule from vacuum to solvent (water) can be divided into two parts, namely, a polar electrostatic effect and a nonpolar cavitation energy. Although the polar component was taken into account in the above analysis of the electrostatic interaction, the nonpolar aspect still remains unexplored. Therefore, we employed an empirical additive model[50] to investigate desolvation effects accompanying protein−DNA binding. In contrast to the favorable contribution of desolvation effects to protein−small ligand,[99] protein−peptide,[100] and protein−protein interactions,[101] desolvation upon binding yields an energetic penalty of 11.47 ± 12.4 kcal·mol^{-1} for protein−DNA association. Apparently, the opposite signs of the desolvation contributions to protein−DNA binding and to other protein-involved binding suggest a significant difference in interfacial properties. Considering that hydrophilicity is a basic attribute of protein−DNA interface, the few protein−DNA complexes with favorable desolvation contributions (manifested by negative ΔG_{desolv} values in Table S5, Supporting Information) are thus particularly interesting because these outliers can reveal some valuable information underlying protein−DNA recognition. The two most noticeable outliers are the TATA-binding protein (TBP) and trimerized small, acid-soluble spore protein (SASP) in complex with their cognate DNA segments, which have ΔG_{desolv} values of −10.49 and −19.67 kcal·mol^{-1}, respectively. As seen in Figure 6, TBP and SASP place β-sheet and α-helix elements, respectively, in DNA minor grooves to achieve close contact with the aromatic bases, but the negatively charged DNA backbones remain untouched by the protein. As a result, the strongly hydrophobic interfaces are defined in the region between the proteins and DNA segments. These two special cases might suggest that the polar character of the protein−DNA interface is conferred by the interactions of the proteins with the DNA backbones, rather than with the DNA bases. In addition, it is noteworthy that the DNA sequence compositions in the two complexes are quite different—one is TA-rich and the

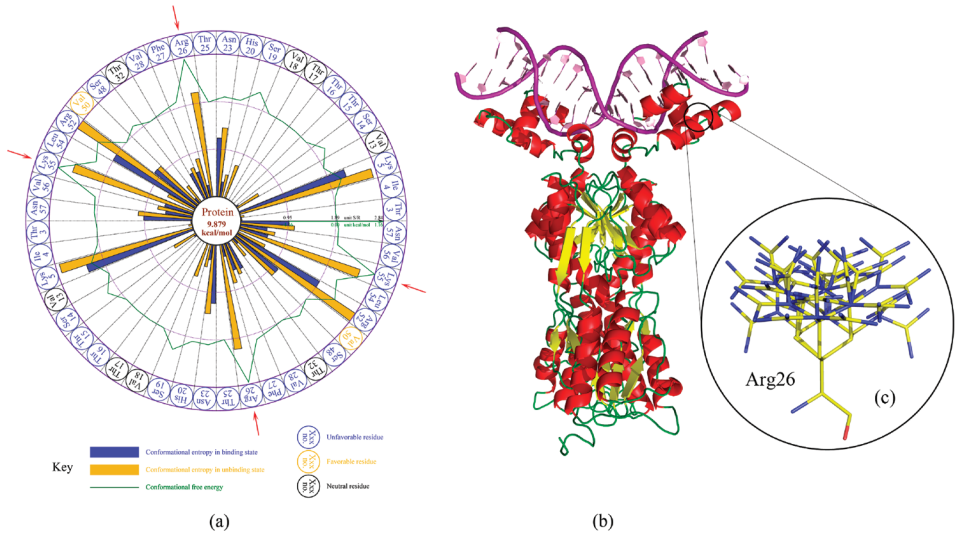

Figure 7. (a) Schematic representation of the side-chain conformational entropy/energy loss involved in interfacial amino acids due to the binding of purine repressor with its cognate DNA. In this diagram, the pie chart is divided equally, with each section representing an interfacial amino acid. Within a sector, side-chain conformational entropies in unbinding and binding states are colored yellow and blue, respectively. The green polygonal line was formed by linking conformational energies of each interfacial amino acid. The conformational entropy and conformational energy for each interfacial amino acid can be measured qualitatively according to the horizontal black and green scales, respectively. In the periphery, amino acid symbols are colored yellow, blue, and black, respectively, to indicate favorable, unfavorable, and neutral contributions to the binding. Four positively charged, long-side-chain amino acids (Arg and Lys each for two) with conformational energy penalties of >1 kcal/mol are marked by red arrows. This diagram was prepared using the in-house program 2D-GraLab.[66] (b) Stereoview of the purine repressor−DNA complex (PDB: 2PUD). (c) Arg26 rotamers.

other is CG-rich—indicating that hydrophobicity is independent of base type.

Conformational Entropy Loss. Compared to the other enthalpy-driven nonbonding effects discussed in the preceding sections, the role of entropy in biological systems is more elusive. Although conformational entropy loss involving protein folding and protein−protein binding have been investigated using both experimental and computational approaches,[102] the conformational entropy effect accompanying protein−DNA association still remains largely unexplored. Normally, the total conformational entropy of a protein can be divided into two parts separately associated with the main chain and the side chain. (DNA entropy was ignored because the DNA helix is quite rigid and thus makes only a small entropy contribution to the binding.) At present, the change in main-chain entropy due to binding can be estimated only at a modest level of accuracy using exhaustive molecular dynamics (MD) normal-mode analysis. Therefore, in this study, we calculated only the side-chain conformational entropy loss using the Boltzmann formula[62] coupled with a well-designed rotamer library.[63] It is well-known that a polypeptide side chain is a dynamic assembly of rotamers.[64] When a protein approaches a locally rigid DNA molecule, a large reduction in the number of side-chain conformations accessible to the rotamer space occurs at the interfacial amino acid residues, leading to a significant loss in their side-chain conformational entropy. According to our analysis, the energetic penalty (conformational energy) arising from entropy loss is as much as 6.70 ± 6.8 kcal·mol^{-1} per protein−DNA complex, which can completely counteract the stabilization energy given by vdW packing or nearly one-half of the desolvation effect. At a global complex level, the conformational energy is closely correlated with the interface area, with a correlation coefficient of $R = 0.817$ for the fitting equation $\Delta G_{SCE} = -0.961 + 0.003B$ (this simple linear model can be used to rapidly estimate the conformational energy for protein−DNA binding) (Figure 5). At a local residue level, the conformational energy is highly dependent on amino acid type; the two positively charged, long-side-chain amino acids Arg and Lys contribute nearly 70% of the total conformational energy to the binding (Figure 7). Interestingly, very few small, nonpolar amino acids, such as Val, exhibit a slight increase in their entropies during binding. This can be ascribed to the averaging of the rotamers' potential by the presence of adjacent but noncollisional DNA atoms.

4. CONCLUSIONS

Traditional structural taxonomy describes protein−DNA recognition as a simple classification of DNA-binding domains. This description derives from the combination of secondary-structure elements at local binding sites and is the current basis for elucidating the structural principles of proteins bound to DNA. However, protein−DNA interactions are too diverse to be described using a simple classification scheme, particularly considering that there exist a large number of structurally known protein−DNA complexes bound in intricate modes, such as the multisubunit enzyme−DNA intertwiners. Instead of the DNA-binding domain, in this work, we systematically investigated the themes in protein−DNA recognition and defined a thematically nonredundant data set. Based on this data set, various physical and chemical properties involved in protein−DNA binding were calculated, analyzed, and discussed in detail. We expect that the newly defined classification tree and nonredundant set should be valuable for further study of the molecular and structural basis of protein−DNA recognition in the context of a uniform framework.

ACKNOWLEDGMENT

We are grateful to the reviewers for their help in improving the technique and language of this article. This work was supported by the State Key Laboratory of Trauma, Burns and Combined Injury Foundation (No. SKLKF200904).

Supporting Information Available: Lists of PDB entries of the 1192 selected and 151 excluded protein−DNA complexes (Tables S1 and S2, respectively); classification results of the themes in protein−DNA recognition (Table S3); detailed descriptions of each trunk, branch, and leaf of the theme classification tree (Table S4); and list of energetic components involved in representative protein−DNA complexes (Table S5). This material is available free of charge via the Internet at http://pubs.acs.org/.

REFERENCES AND NOTES

(1) Sarai, A.; Kono, H. Protein−DNA recognition patterns and predictions. *Annu. Rev. Biophys. Biomol. Struct.* **2005**, *34*, 379–398.
(2) Matthews, B. W. Protein−DNA interaction. No code for recognition. *Nature* **1988**, *335*, 294–295.
(3) Heinemann, U.; Büssow, K.; Mueller, U.; Umbach, P. Facilities and methods for the high-throughput crystal structural analysis of human proteins. *Acc. Chem. Res.* **2003**, *36*, 157–163.
(4) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The protein data bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
(5) Levitt, M.; Chothia, C. Structural patterns in globular proteins. *Nature* **1976**, *261*, 552–557.
(6) Chothia, C.; Levitt, M.; Richardson, D. Structure of proteins: Packing of α-helices and β-sheets. *Proc. Natl. Acad. Sci. U.S.A.* **1977**, *74*, 4130–4134.
(7) Chothia, C. Principles that determine the structure of proteins. *Annu. Rev. Biochem.* **1984**, *53*, 537–572.
(8) Orengo, C. A.; Flores, T. P.; Taylor, W. R.; Thornton, J. M. Recurring structural motifs in proteins with different functions. *Curr. Biol.* **1993**, *3*, 131–139.
(9) Flores, T. P.; Orengo, C. A.; Thornton, J. M. Conformational characteristics in structurally similar protein pairs. *Protein Sci.* **1993**, *7*, 31–37.
(10) Orengo, C. A.; Jones, D. T.; Taylor, W.; Thornton, J. M. Protein superfamilies and domain superfolds. *Nature* **1994**, *372*, 631–634.
(11) Richardson, J. S. β-sheet topology and the relatedness of proteins. *Nature* **1977**, *268*, 495–500.
(12) Richardson, J. S. The anatomy and taxonomy of protein structure. *Adv. Protein Chem.* **1981**, *34*, 167–339.
(13) Richardson, J. S.; Richardson, D. C. The *de novo* design of protein structures. *Trends Biochem. Sci.* **1989**, *14*, 304–309.
(14) Murzin, A. G.; Brenner, S. E.; Hubbard, T.; Chothia, C. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **1995**, *247*, 536–540.
(15) Orengo, C. A.; Michie, A. D.; Jones, D. T.; Swindells, M. B.; Thornton, J. M. CATH−A hierarchic classification of protein domain structures. *Structure* **1997**, *5*, 1093–1108.
(16) Klosterman, P. S.; Tamura, M.; Holbrook, S. R.; Brenner, S. E. SCOR: A structural classification of RNA database. *Nucleic Acids Res.* **2002**, *30*, 392–394.
(17) Winter, C.; Henschel, A.; Kim, W. K.; Schroeder, M. SCOPPI: A structural classification of protein−protein interfaces. *Nucleic Acids Res.* **2006**, *34*, D310−D314.
(18) Teyra, J.; Paszkowski-Rogacz, M.; Anders, G.; Pisabarro, M. T. SCOWLP classification: Structural comparison and analysis of protein binding regions. *BMC Bioinf.* **2008**, *9*, 9.

(19) Pabo, C. O.; Nekludova, L. Geometric analysis and comparison of protein−DNA interfaces: Why is there no simple code for recognition. *J. Mol. Biol.* **2000**, *301*, 597–624.
(20) Kim, W. K.; Henschel, A.; Winter, C.; Schroeder, M. The many faces of protein−protein interactions: A compendium of interface geometry. *PLoS Comput. Biol.* **2006**, *2*, e124.
(21) Ahmad, S.; Keskin, O. S.; Sarai, A.; Nussinov, R. Protein−DNA interactions: Structural, thermodynamic and clustering patterns of conserved residues in DNA-binding proteins. *Nucleic Acids Res.* **2008**, *36*, 5922–5932.
(22) Sutch, B. T.; Chambers, E. J.; Bayramyan, M. Z.; Gallaher, T. K.; Haworth, I. S. Similarity of protein−RNA interfaces based on motif analysis. *J. Chem. Inf. Model.* **2009**, *49*, 2139–2146.
(23) Harrison, S. C. A structural taxonomy of DNA-binding domains. *Nature* **1991**, *353*, 715–719.
(24) Luisi, B. F. DNA−protein interaction at high resolution. In *DNA−Protein Structural Interactions*; Lilley, D. M. J., Ed.; Oxford University Press: New York, 1995; pp 1−48.
(25) Luscombe, N. M.; Austin, S. E.; Berman, H. M.; Thornton, J. M. An overview of the structures of protein−DNA complexes. *Genome Biol.* **2000**, *1*, 1–37.
(26) Struhl, K. Helix-turn-helix, zinc-finger, and leucine-zipper motifs for eukaryotic transcriptional regulatory proteins. *Trends Biochem. Sci.* **1989**, *14*, 137–140.
(27) Prabakaran, P.; Ahmad, S.; Gromiha, M. M.; Singarayan, M. G.; Sarai, A. Classification of protein−DNA complexes based on structural descriptors. *Structure* **2006**, *14*, 1355–1367.
(28) Siggers, T. W.; Silkov, A.; Honig, B. Structural alignment of protein−DNA interfaces: Insights into the determinants of binding specificity. *J. Mol. Biol.* **2005**, *345*, 1027–1045.
(29) Kono, H.; Sarai, A. Structure-based prediction of DNA target sites by regulatory proteins. *Proteins* **1999**, *35*, 114–131.
(30) von Hippel, P. H. Protein−DNA recognition: New perspectives and underlying themes. *Science* **1994**, *263*, 769–770.
(31) Draper, D. E. Themes in RNA−protein recognition. *J. Mol. Biol.* **1999**, *293*, 255–270.
(32) Zhou, P.; Zou, J.; Tian, F.; Shang, Z. Geometric similarity between protein−RNA interfaces. *J. Comput. Chem.* **2009**, *30*, 2738–2751.
(33) Jones, S.; van Heyningen, P.; Berman, H. M.; Thornton, J. M. Protein−DNA interactions: A structural analysis. *J. Mol. Biol.* **1999**, *287*, 877–896.
(34) Berman, H. M.; Olson, W. K.; Beveridge, D. L.; Westbrook, J.; Gelbin, A.; Demeny, T.; Hsieh, S.-H.; Srinivasan, A. R.; Schneider, B. The nucleic acid database: A comprehensive relational database of three-dimensional structures of nucleic acids. *Biophys. J.* **1992**, *63*, 751–759.
(35) Word, J. M.; Lovell, S. C.; Richardson, J. S.; Richardson, D. C. Asparagine and glutamine: Using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.* **1999**, *285*, 1735–1747.
(36) Krivov, G. G.; Shapovalov, M. V.; Dunbrack, R. L., Jr. Improved prediction of protein side-chain conformations with SCWRL4. *Proteins* **2009**, *77*, 778–795.
(37) McDonald, I. K.; Thornton, J. M. Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.* **1994**, *238*, 777–793.
(38) Word, J. M.; Lovell, S. C.; LaBean, T. H.; Taylor, H. C.; Zalis, M. E.; Presley, B. K.; Richardson, J. S.; Richardson, D. C. Visualizing and quantifying molecular goodness-of-fit: Small-probe contact dots with explicit hydrogen atoms. *J. Mol. Biol.* **1999**, *285*, 1711–1733.
(39) Carugo, O.; Argos, P. Protein−protein crystal-packing contacts. *Protein Sci.* **1997**, *6*, 2261–2263.
(40) Phipps, K. R.; Li, H. Protein−RNA contacts at crystal packing surfaces. *Proteins* **2007**, *67*, 121–127.
(41) Kabsch, W.; Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577–2637.
(42) Kim, W. K.; Ison, J. C. Survey of the geometric association of domain-domain interfaces. *Proteins* **2005**, *61*, 1075–1088.
(43) Tuncbag, N.; Gursoy, A.; Guney, E.; Nussinov, R.; Keskin, O. Architectures and functional coverage of protein−protein interfaces. *J. Mol. Biol.* **2008**, *381*, 785–802.
(44) Smith, T. F.; Waterman, M. S. Identification of common molecular subsequences. *J. Mol. Biol.* **1981**, *147*, 195–197.
(45) Jeong, E.; Kim, H.; Lee, S.-W.; Han, K. Discovering the interaction propensities of amino acids and nucleotides from protein−RNA complexes. *Mol. Cells* **2007**, *16*, 161–167.
(46) 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.
(47) Donald, J. E.; Chen, W. W.; Shakhnovich, E. I. Energetics of protein−DNA interactions. *Nucleic Acids Res.* **2007**, *35*, 1039–1047.
(48) Boobbyer, D. N. A.; Goodford, P. J.; McWhinnie, P. M.; Wade, R. C. New hydrogen-bond potentials for use in determining energetically favorable binding sites on molecules of known structure. *J. Med. Chem.* **1989**, *32*, 1083–1094.
(49) Hagler, A. T.; Huler, E.; Lifson, S. Energy functions for peptides and proteins. I. Derivation of a consistent force field including the hydrogen bond from amide crystals. *J. Am. Chem. Soc.* **1974**, *96*, 5319–5327.
(50) Fraternali, F.; Cavallo, L. Parameter optimized surfaces (POPS): Analysis of key interactions and conformational changes in the ribosome. *Nucleic Acids Res.* **2002**, *30*, 2950–2960.
(51) Fraternali, F.; van Gunsteren, W. F. An efficient mean solvation force model for use in molecular dynamics simulations of proteins in aqueous solution. *J. Mol. Biol.* **1996**, *256*, 939–948.
(52) Sanner, M. F.; Olson, A. J.; Spehner, J.-C. Reduced surface: An efficient way to compute molecular surfaces. *Biopolymers* **1996**, *38*, 305–320.
(53) Tsai, J.; Taylor, R.; Chothia, C.; Gerstein, M. The packing density in proteins: Standard radii and volumes. *J. Mol. Biol.* **1999**, *290*, 253–266.
(54) Nadassy, K.; Tomás-Oliveira, I.; Alberts, I.; Janin, J.; Wodak, S. J. Standard atomic volumes in double-stranded DNA and packing in protein−DNA interfaces. *Nucleic Acids Res.* **2001**, *29*, 3362–3376.
(55) Zhou, H.-X. Macromolecular electrostatic energy within the nonlinear Poisson-Boltzmann equation. *J. Chem. Phys.* **1994**, *100*, 3152–3162.
(56) Rocchia, W.; Alexov, E.; Honig, B. Extending the applicability of the nonlinear. Poisson−Boltzmann equation: Multiple dielectric constants and multivalent ions. *J. Phys. Chem.* **2001**, *105*, 6507–6514.
(57) Yan, S.; Wu, M.; Pate, D. J.; Geacintov, N. E.; Broyde, S. Simulating structural and thermodynamic properties of carcinogen-damaged DNA. *Biophys. J.* **2003**, *84*, 2137–2148.
(58) Gilson, M. K.; Honig, B. Calculation of electrostatic potentials in an enzyme active site. *Nature* **1987**, *330*, 84–86.
(59) Qin, S.; Zhou, H.-X. Do electrostatic interactions destabilize protein−nucleic acid binding. *Biopolymers* **2007**, *86*, 112–118.
(60) Sitkoff, D.; Sharp, K. A.; Honig, B. Accurate calculation of hydroation free energies using macroscopic solvent models. *J. Phys. Chem.* **1994**, *98*, 1978–1988.
(61) Frederick, K. K.; Marlow, M. S.; Valentine, K. G.; Wand, A. J. Conformational entropy in molecular recognition by proteins. *Nature* **2007**, *448*, 325–330.
(62) Zhang, J.; Liu, J. S. On side-chain conformational entropy of proteins. *PLoS Comput. Biol.* **2006**, *2*, e168.
(63) Lovell, S. C.; Word, J. M.; Richardson, J. S.; Richardson, D. C. The penultimate rotamer library. *Proteins* **2000**, *40*, 389–408.
(64) Dunbrack, R. L., Jr. Rotamer libraries in the 21st century. *Curr. Opin. Struct. Biol.* **2002**, *12*, 431–440.
(65) Koehl, P.; Delarue, M. Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. *J. Mol. Biol.* **1994**, *239*, 249–275.
(66) Zhou, P.; Tian, F.; Shang, Z. 2D depiction of nonbonding interactions for protein complexes. *J. Comput. Chem.* **2009**, *30*, 940–951.
(67) Nadassy, K.; Wodak, S. J.; Janin, J. Structural features of protein−nucleic acid recognition sites. *Biochemistry* **1999**, *38*, 1999–2017.
(68) Orengo, C. A.; Taylor, W. R. SSAP: Sequential structure alignment program for protein structure comparison. *Methods Enzymol.* **1996**, *266*, 617–635.
(69) McLachlan, A. D. Rapid comparison of protein structures. *Acta Crystallogr. A* **1982**, *38*, 871–873.
(70) Frishman, D.; Mewes, H.-W. PEDANTic genome analysis. *Trends Genet.* **1997**, *13*, 415–416.
(71) Tsai, C. J.; Lin, S. L.; Wolfson, H. J.; Nussinov, R. A dataset of protein−protein interfaces generated with a sequence-order-independent comparison technique. *J. Mol. Biol.* **1996**, *260*, 604–620.
(72) Keskin, O.; Tsai, C.-J.; Wolfson, H.; Nussinov, R. A new, structurally nonredundant, diverse data set of protein−protein interfaces and its implications. *Protein Sci.* **2004**, *13*, 1043–1055.
(73) Keskin, O.; Gursoy, A.; Ma, B.; Nussinov, R. Principles of protein−protein interactions: What are the preferred ways for proteins to interact. *Chem. Rev.* **2008**, *108*, 1225–1244.
(74) Mintz, S.; Shulman-Peleg, A.; Wolfson, H. J.; Nussinov, R. Generation and analysis of a protein−protein interface data set with similar chemical and spatial patterns of interactions. *Proteins* **2005**, *61*, 6–20.
(75) Chothia, C. One thousand families for the molecular biologist. *Nature* **1992**, *357*, 543–544.
(76) Aloy, P.; Russell, R. B. Ten thousand interactions for the molecular biologist. *Nat. Biotechnol.* **2004**, *22*, 1317–1321.
(77) Aloy, P.; Russell, R. B. Interrogating protein interaction networks through structural biology. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 5896–5901.

(78) Aloy, P.; Russell, R. B. The third dimension for protein interactions and complexes. *Trends Biochem. Sci.* **2002**, *27*, 633–638.
(79) Aloy, P.; Russell, R. B. InterPreTS: Protein interaction prediction through tertiary structure. *Bioinformatics* **2003**, *19*, 161–162.
(80) Davis, F. P.; Sali, A. The overlap of small molecule and protein binding sites within families of protein structures. *PLoS Comput. Biol.* **2010**, *6*, e1000668.
(81) 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.
(82) Lejeune, D.; Delsaux, N.; Charloteaux, B.; Thomas, A.; Brasseur, R. Protein−nucleic acid recognition: Statistical analysis of atomic interactions and influence of DNA structure. *Proteins* **2005**, *61*, 258–271.
(83) Kim, O. T. P.; Yura, K.; Go, N. Amino acid residue doublet propensity in the protein−RNA interface and its application to RNA interface prediction. *Nucleic Acids Res.* **2006**, *34*, 6450–6460.
(84) Ellis, J. J.; Broom, M.; Jones, S. Protein−RNA interactions: Structural analysis and functional classes. *Proteins* **2007**, *66*, 903–911.
(85) Gromiha, M. M.; Siebers, J. G.; Selvaraj, S.; Kono, H.; Sarai, A. Intermolecular and intramolecular readout mechanisms in protein−DNA recognition. *J. Mol. Biol.* **2004**, *337*, 285–294.
(86) Ponnuswamy, P. K.; Gromiha, M. M. On the conformational stability of folded proteins. *J. Theor. Biol.* **1994**, *166*, 63–74.
(87) Ponnuswamy, P. K.; Gromiha, M. M. On the conformational stability of oligonucleotide duplexes and tRNA molecules. *J. Theor. Biol.* **1994**, *169*, 419–432.
(88) Lo Conte, L.; Chothia, C.; Janin, J. The atomic structure of protein−protein recognition sites. *J. Mol. Biol.* **1999**, *285*, 2177–2198.
(89) Tsai, C. J.; Nussinov, R. Hydrophobic folding units at protein−protein interfaces: Implication to protein folding and to protein−protein association. *Protein Sci.* **1997**, *6*, 1426–1437.
(90) Tomovic, A.; Oakeley, E. J. Computational structural analysis: Multiple proteins bound to DNA. *PLoS ONE* **2008**, *3*, e3243.
(91) Bello, J. Tight packing of protein cores and interfaces. *Int. J. Pept. Protein Res.* **1978**, *12*, 38–41.
(92) Roth, C. M.; Neal, B. L.; Lenhoff, A. M. Van der Waals interactions involving proteins. *Biophys. J.* **1996**, *70*, 977–987.
(93) Boger, D. L.; Invergo, B. J.; Coleman, R. S.; Zarrinmayeh, H.; Kitos, P. A.; Thompson, S. C.; Leong, T.; McLaughlin, L. W. A demonstration of the intrinsic importance of stabilizing hydrophobic binding and non-covalent van der Waals contacts dominant in the non-covalent CC-1065/B-DNA binding. *Chem. Biol. Interact.* **1990**, *73*, 29–52.
(94) Rother, K.; Hildebrand, P. W.; Goede, A.; Gruening, B.; Preissner, R. Voronoia: Analyzing packing in protein structures. *Nucleic Acids Res.* **2009**, *37*, D393–D935.
(95) Li, A.-J.; Nussinov, R. A set of van der Waals and Coulombic radii of protein atoms for molecular and solvent-accessible surface calculation, packing evaluation, and docking. *Proteins* **1998**, *32*, 111–127.
(96) Misra, V. K.; Hecht, J. L.; Yang, A. S.; Honig, B. Electrostatic contributions to the binding free energy of the lambda cl repressor to DNA. *Biophys. J.* **1998**, *75*, 2262–2273.
(97) Olson, M. A. Calculations of free-energy contributions to protein−RNA complex stabilization. *Biophys. J.* **2001**, *81*, 1841–1853.
(98) Reyes, C. M.; Kollman, P. A. Structure and thermodynamics of RNA−protein binding: Using molecular dynamics and free energy analyses to calculate the free energies of binding and conformational change. *J. Mol. Biol.* **2000**, *297*, 1145–1158.
(99) Kenji, A.; Yasuo, N. Hydrophobic effect on the protein−ligand interaction. *Chem. Pharm. Bull.* **1989**, *37*, 86–92.
(100) Cserhati, T.; Szogyi, M. Role of hydrophobic and hydrophilic forces in peptide−protein interaction: New advances. *Peptides* **1995**, *16*, 165–173.
(101) Jones, S.; Thornto, J. M. Principles of protein−protein interactions. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 13–20.
(102) Brady, G. P.; Sharp, K. A. Entropy in protein folding and in protein−protein interactions. *Curr. Opin. Struct. Biol.* **1997**, *7*, 215–221, and references therein.

CI100145D